Protein Structure Prediction Enhanced with Evolutionary Diversity: SPEED

Joe DeBartolo[1], Glen Hocky[2], Mike Wilde[4,7], Jinbo Xu[6], Karl F. Freed[2,3,4,5], and Tobin R. Sosnick[1,4,5]

[1]Department of Biochemistry and Molecular Biology, Univ. of Chicago
[2]Department of Chemistry, Univ. of Chicago
[3]The James Franck Institute, Univ. of Chicago
[4]Computation Institute, Univ. of Chicago
[5]Institute for Biophysical Dynamics, Univ. of Chicago
[6]Toyota Technological Institute at Chicago
[7]Argonne National Laboratory

Corresponding authors:

Tobin R. Sosnick: trsosnic@uchicago.edu, 773-218-5950

Karl F. Freed: freed@uchicago.edu,

Abstract. For naturally occurring proteins, similar sequence implies similar structure. Consequently, multiple sequence alignments often are used in template-based modeling of protein structure and have been incorporated into fragment-based assembly methods. Our previous homology-free structure prediction study introduced an algorithm that mimics the folding pathway by coupling the formation of secondary and tertiary structure. Moves in the Monte Carlo procedure involve only a change in a single pair of ϕ,ψ backbone dihedral angles that are obtained from a PDB-based distribution appropriate for each amino acid, conditional on the type and conformation of the flanking residues. We improve this method by utilizing multiple sequence alignments to enrich the sampling distribution, but in a manner that does not require structural knowledge of any protein sequence (i.e., not fragment insertion). In combination with other tools, including clustering and refinement, the accuracies of the predicted secondary and tertiary structures are substantially improved and a global and position-resolved measure of confidence is introduced for the accuracy of the predictions. Performance of the method in the Critical Assessment of Structure Prediction (CASP8) is discussed.

Keywords: Protein folding, multiple sequence alignment, ItFix, folding pathway, statistical potential, Monte Carlo simulated annealing

Abbreviations. ItFix, iterative fixing; MCSA, Monte Carlo Simulated Annealing; MSA, multiple sequence alignment; 2^o structure, secondary structure; 3^o structure, tertiary structure.

Introduction

Given the expansion of the sequence database, an imperative of the field of structural biology is to cluster related sequences into families and determine a representative structure for each family(Fitch 1970; Murzin et al. 1995; Li et al. 2001; Bateman et al. 2002; Levitt 2009). The already large number of families is rapidly expanding and the cost of determining representative protein structures is high. Computational structure prediction may provide the most effective means of mapping the protein universe. Structure prediction, however, is inherently challenging because of the enormous conformational space accessible to each amino acid sequence. For this reason, the most successful prediction methods seek to narrow the conformational search, for example by using large PDB fragments (Raman et al. 2009) rather than simulating the protein *ab initio(Ozkan et al. 2007; Yang et al. 2007)*.

We have recently developed a C_β-level, homology-free structure prediction algorithm, termed ItFix,(DeBartolo et al. 2009) in which the conformational search space is restricted by iteratively fixing secondary (2°) structure assignments of certain portions of the sequence after incorporating the influence of tertiary (3°) context. Moreover, the iterative feature enables regions of lower confidence to be predicted after the fixing of more confident regions. The coupling and mutual stabilization of 2° and 3° structure formation mimics the pathway character exhibited by real proteins(Krantz et al. 2004; Sosnick et al. 2006).

The computationally rapid, homology-free algorithm uses moves involving only the change in a single pair of ϕ,ψ dihedral angles (pivot moves). Hence, its performance is independent of the existence of appropriate fragments from the PDB. Nevertheless, our algorithm can outperform current homology-based 2° structure prediction methods for many

proteins. ItFix also generates 3° structures of comparable accuracy to existing methods for many small proteins, including ones with few sequence homologues.

Our earlier study revealed that a large impediment to more accurate structure prediction arises from the intrinsically low propensity of some residues to adopt the backbone dihedral angles found in their native structures. In the protein 1dcj, for example, the middle of a helix contains a proline followed by a glycine, two residues that are very unlikely to be found together in helices. Even though ItFix uses more confidently assigned regions to identify native structure in otherwise weakly determined regions, the additional contextual information occasionally is insufficient to override very strong local biases. Unfortunately, issues of this severity occur often in many proteins, and the associated errors can detrimentally affect the accuracy of the 2° and 3° structure prediction.

Here, we employ multiple sequence alignments (MSAs) to mitigate the influence of the non-native local biases. MSAs are incorporated into many popular 2° structure(Jones 1999; Pollastri et al. 2002) and both template-based(Skolnick et al. 2000; Baker and Sali 2001; Zhou and Skolnick 2009) and template-free(Bradley et al. 2005; Zhao et al. 2008) 3° structure prediction methods. In our distribution of sampled ϕ,ψ angles, the non-native biases are manifested as a low probability of native-like angles. This PDB-based distribution is now enriched using the sequence diversity found in an MSA, but does so *without requiring structural information from any constituent sequence*. We denote this procedure SPEED: Structure Prediction Enhanced by Evolutionary Diversity (Fig. 1). The combination of ItFix and SPEED significantly increases the accuracy of 2° and 3° structure predictions, and more so in combination with novel energy functions and clustering methods. We also provide global and

local measures of the confidence of our predictions, thereby providing an essential tool for assessing the accuracy of the predicted structures of unsolved sequence families.

Results

Overview. Figure 1a provides an overview of both the homology-free and SPEED structure prediction methods utilizing the ItFix 2° structure fixing procedure. The fundamental difference between our original homology-free protocol and the new SPEED protocol relates to the Ramachandran (Rama) ϕ,ψ sampling distribution. In the homology-free protocol, the distribution is generated only from the target sequence, whereas in the new protocol, the distribution is constructed from an MSA of the target sequence. At the beginning of the ItFix procedure, no 2° structure is fixed, and the ϕ,ψ distribution at each position reflects all 2° structure types, although the distribution is contingent on the amino acid identities of the neighboring positions (Fig. 1b). Through rounds of folding (Monte Carlo simulated annealing, MCSA) using an energy function that promotes hydrophobic burial and that penalizes polar burial (Methods), the 2° structure options helix, strand or coil are progressively eliminated when their occurrence in the final collapsed structures falls below a ~0-10% threshold(DeBartolo et al. 2009). Angles originating from the eliminated 2° structure option are excluded in the calculation of the Rama distribution for the subsequent round. The folding and elimination process proceeds until no further 2° structure options can be eliminated (Fig. 1b middle and bottom). The final result is a more restricted Rama distribution across the entire sequence which greatly reduces the search space.

The final Rama distribution is used to generate a large (10,000) ensemble of 3° structure models. These models are clustered into groups of similar structure, and the models from the

largest cluster are selected for refinement and prediction, using our DOPE-PW statistical potential.

SPEED enhanced Ramachandran distributions. At the beginning of the ItFix rounds, the Rama distribution at each position is conditional only on the amino acid identities of the position and its two neighbors. Our homology-free implementation obtains this distribution solely using the target sequence. For example, N4 of 1tif is flanked by I3 and E5 (denoted $_{I}N_{E}$), with the resulting INE having a homology-free Rama distribution displayed in the left panel of Figure 1b. The SPEED-enhanced Rama distribution is the sum (with equal weights) of the distributions of all possible three-residue combinations generated from the amino acid substitutions identified by the MSA. For example, the SPEED distribution for $_{I}N_{E}$ is the sum of multiple Rama distributions derived from the MSA, including $_{I}N_{D}$, $_{I}G_{D}$, $_{V}G_{N}$, etc. At the beginning of the algorithm when no 2° structure option is eliminated, the native Rama region (Fig. 1b, red circle) has a small sampling probability in the homology-free distribution (P=0.01), and the predominant Rama region is right-handed helix (P=0.6). By contrast, the native Rama region has a ~20-fold larger probability in the equivalent SPEED Rama distribution. Also, at the end of the ItFix rounds, the SPEED probability of the native Rama region has nearly doubled compared to the homology-free probability (P=0.37 versus 0.21). The native Rama probability enhancement due to ItFix is thus significantly improved by MSA-based procedure.

To illustrate the benefit of using SPEED, we quantify the enhancement across all positions in the folding targets by comparing the native Rama probability of the homology-free distribution to that of the SPEED-derived distribution (Fig. 2). This analysis proceeds by partitioning the Rama map into four broad regions (Fig. 2a). More refined divisions of the Rama

map exist, but this division into four regions may be the most refined definition with clear borders. The quality of SPEED-derived distribution is quantified as the percentage of positions with low probability of the native Rama region (P<0.25). This percentage is a useful metric because any position with such a low native Rama probability is an obvious candidate for improvement. Compared to the homology-free Rama distributions, the new procedure decreases the percentage of residues having a non-native Rama propensity for 10 out of the 12 targets studied (Fig. 2b). The two exceptions remain unchanged because their homology-free distributions already are very good. The two targets with the largest improvement in Rama distribution are 1csp (78% → 86%) and 1dcj (84% → 94%). In particular, the homology-free Rama distribution for 1dcj contains serious flaws due to the aforementioned proline-glycine pair in the second α-helix and for residues in the turn separating the second helix and third strand (Suppl. Fig. S1a). SPEED overrides the non-native propensity of G46 in the second helix (P=0.21 → P=0.62) and also enhances the E52 turn position's native propensity (P=0.01 → P=0.32).

In addition to the moderation of outliers, SPEED enhances the native Rama propensity when it is already high, as is the case for 1b72. Here, the native Rama probability at only one of the ten coil positions (E31) falls below the 0.25 threshold (Fig. 2c). Its native-like probability is only P=0.03 in the homology-free distribution but is enhanced to P=0.23 in the enhanced distribution. Additionally, the native Rama probability in the SPEED-derived distribution is two-fold higher than the homology-free distribution in 7 out of 10 coil positions. Similar improvements for other targets can be seen in Supplemental Figure 1.

The exceptions to this trend generally emerge for positions which already have a very strong native-like propensity in the target sequence. An illustration of this effect is the left-

handed turn position G10 in 1ubq. Because glycine favors the native left-handed turn basin more than all other residues, any substitution lowers the native Rama probability (Suppl. Fig. S1b). Nevertheless, the decrease in native probability due to the use of SPEED is on average is much smaller than the benefit at other positions (Fig. 2, Suppl. Fig. S1).

ItFix 2° structure. The 2° structures of the final models are identified using the DSSP program for 2° structure determination(Kabsch and Sander 1983). Since DSSP-identified β-strands must be involved in β-sheet networks with optimized hydrogen bonds, the strand-fixing threshold is lower than our previous study, with no noticeable decrease in fidelity. In many cases, the fidelity for specifying 2° structure is higher. This increase is particularly evident for the all-α targets, where the β-strand option is eliminated at every position after the first two rounds where the β-strand probability very nearly vanishes ($P< 0.005$) at every position (1af7, 1b72 in the first round; 1r69 in the second round). The same accuracy is found for the helical regions of the αβ targets.

Improvement in 2° prediction accuracy. The 2° structure prediction accuracy using SPEED compares very favorably with the popular 2° structure prediction methods SSPro(Pollastri et al. 2002) and PSIPRED(Jones 1999) (Table 1). When predicting 2° structure at the level of helix, extended or coil (three options, termed Q3), ItFix-SPEED is more accurate than its homology-free ItFix counterpart (average accuracy 84% → 88%). Most of this improvement is due to 1csp (79% → 87%) and 1dcj (45% → 83%), the two targets with the largest improvements in Rama distribution due to SPEED (Fig. 2b). The 2° structures for the all-α targets already are predicted to high accuracy using the homology-free ItFix, so the average improvement due to SPEED is

small (93% → 96%), with the exception of 1b72 where the improvement is more substantial (88% → 96%).

More impressive is the increase in accuracy for the prediction of 2° structure at the more refined Q8 level where coil is subdivided into six DSSP-identified subtypes (this level of prediction is unavailable with PSIPRED). For 1b72, the overall Q8 accuracy increases (84% → 96%) using SPEED with a >0.95 probability assigned to the native Q8 value at every position in the second coil region. Two other targets that have substantial improvements in Q8 accuracy are 1dcj (29% → 65%) and 1ubq (69% → 82%). Most of the Q8 improvements for 1dcj arise from the same helix and strand improvements found for the Q3 values, whereas the Q8 improvements for 1ubq are due almost exclusively to better turn predictions within the coil subtype.

Energy Functions. We continue to use a reduced C_β model that includes the backbone heavy atoms, backbone amide hydrogen, and the side chain C_β, and a slightly modified version of the DOPE-PW energy function(DeBartolo et al. 2009). This energy function is a pairwise additive statistical potential based on the observed distance distributions between each atom in the model. In addition to distinguishing each type of atom, the energy function classifies each interaction according to residue type, 2^o structure, and side-chain orientation.

In the prior ItFix treatment, the 2° structure assignment used for the energy function calculation is that in the original PDB structure from which the φ,ψ pair is selected. Here, the 2° structure is specified using a geometric definition of 2°structure that is applied in each energy calculation (i.e., in the application of the strand-strand terms, helix-helix terms, etc.). A residue is considered to lie in a helix if it is situated in a block of more than four residues in a row satisfying the following criteria:

$-90 < \phi < -40 \vee -60 < \psi < -20$ and $E_{hb} < -4$,

The minimum distance between the hydrogen bond donors and acceptors is described by the distance criterion from the hydrogen bond potential of Kortemme et al.,(Kortemme et al. 2003)

$$\{ [1.7 < \text{dist}(CO_i, NH_j) < 2.6] \text{ or } [1.7 < \text{dist}(NH_i, CO_j) < 2.6] \}$$

In addition to this distance constraint, the hydrogen bond energy function also considers the influence of hydrogen bond orientation. The following term is used to describe the orientation between two covalent bonds, an example being the backbone carbonyl (C=O) bond and amide bond (N-H) orientation:

$$\rho = (\rho_{12} - 90)^2 + (\rho_{21} - 90)^2,$$

In this equation, ρ_{12} represents the angle between the $\vec{CO}$ and $\vec{NH}$ vectors and ρ_{21} represents the angle between the $\vec{NH}$ and $\vec{CO}$ vectors. We impose a 90° minimum on ρ to maintain a planar sheet network for both parallel and anti-parallel sheet networks.

Our previous study(DeBartolo et al. 2009) finds that the statistical potential alone often is incapable of generating a large proportion of well-collapsed models for the targets that contains β-sheets. These simulation models commonly contain attributes that are uncharacteristic of experimental models, such as buried polar residues, unpaired buried β strands, and a high radius of gyration of C_α atoms (*Rg*). Buried polar residues and buried unpaired beta strands are symptomatic of an energetic benefit allotted for the close pairing of non-polar C_β atoms and the lack of penalty for the close pairing of polar and non-polar C_β atoms. Thus, the prior treatment allows a strand to be buried in the hydrophobic core of a model so long as it contains a sufficient number of non-polar residues. High-*Rg* models can be low in energy due to highly optimized

sub-structures, such as β hairpins, which are formed at the expense of integrating the entire chain into a properly-collapsed model.

Adding a penalty for the burial of polar residues impedes the generation of low-*Rg* models, and forcing a lower *Rg* on the chain can worsen the burial of polar groups and beta-strands. For this reason, in addition to *Rg*, two radial terms are included to encourage the proper global collapse of the entire chain. Radial uniformity (*Ru*) is the standard deviation of the distances of Cα atoms from the Cα center of mass (cm),

$$Ru = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}, \text{ where } d_i = r_i - r_{cm}, \text{ and } \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

The *Ru* term is necessary because small globular single-domain proteins rarely have a completely buried segment of chain, but instead have an amphipathic alternation between exposed and buried side chains. Enforcing a small value of *Ru* prevents any portion of the chain from being too close to the center of mass and therefore diminishes the propensity for the burial of entire 2° structure units in the core of the model.

Rg and *Ru* are minimized to create a collapsed chain with no completely buried chain segments. A third radial term, the ratio of the *Rg* of the non-polar C_β atoms to the *Rg* of the polar Cβ atoms, is called burial ratio (*Br*):

$$Br = Rg_{\text{non-polar}} / Rg_{\text{polar}}$$

Most small proteins have the non-polar C_β atoms closer to the center of the protein, whereas the polar C_β atoms are more likely to be on the exterior, so *Br* is less than unity to capture the global hydrophobic burial of globular proteins. The global burial induced by the *Br* term contrasts to the local optimization of statistical potentials, which can optimize local subsets of hydrophobic atom pairs at the expense of global burial.

We add the three radial terms to obtain the overall scoring function, where $E_{DOPE\text{-}repulsive}$ is sum of the positive (repulsive) DOPE terms,

$$\text{DOPE-PW} = 100 * Ru * Ru + E_{DOPE\text{-}repulsive}$$

Each MCSA simulation is repeated using E_{radial} until the B_r is less than 0.80. We cap the minimum value of *Ru* at 2.5 Å, since it is very easy for the chain to fold into a ring structure with *Ru* close to 0. The multiplied radial terms have a coefficient of 100, so that their combined magnitude is significant relative to the repulsive part of DOPE.

The radial terms are used throughout the ItFix algorithm until the 2° structure is determined. For the final round of folding (10,000 models), if the 2° structure is all-α the DOPE-PW energy function is used, otherwise the DOPE energy function is used. The final model refinement process uses the DOPE-PW energy function for all targets.

Improvement in 3° structure. SPEED significantly improves the quality of 3° models compared to the homology-free treatment (Table 2). The model with the lowest C_α-RMSD (best model) is lower for SPEED in every case except 1di2. Because the best model is not always a very reproducible metric of over-all performance, we consider instead the fraction of final structures below 5 Å C_α-RMSD to the native structure (Fig. 3). This fraction is on average several times greater for SPEED than from the homology-free approach when all other folding parameters (2° structure assignment, energy weighting coefficients, etc.) are identical (Table 2, last column). The SPEED folding ensemble for 1ubq is the most enhanced, containing six times more native-like models than the homology-free ensemble. For four out of the twelve targets, the homology-free distribution produces no models below 5 Å, and hence the SPEED enhancement factor is effectively infinite. Even so, improvement also is evident across all ranges

of C_α-RMSD. For 1b72, the addition of SPEED improves the 3° structure ensemble such that 83% of the models are less than 5 Å Cα-RMSD to the native structure (Fig. 3b), which compares favorably to 76% of the homology-free models falling below that threshold. Compared to the β and αβ targets, the three α targets have the most native-like ensembles for both homology-free and SPEED methods, and, hence, this class yields the smallest enhancement factor. Conversely, the β and αβ targets produce a very small fraction of native-like models for both SPEED and homology-free methods, but have the largest increase in native-like models due to the use of SPEED (Table 2). Neither the SPEED nor homology-free methods generate native-like models for 1di2, most likely because it is considerably more prolate in shape than the rest of the proteins, and the radial energy terms (R*u*, R*g*, B*r* see Methods) enforce a spherical bias (Suppl. Table S1).

Averaging the energy function across the MSA. Analogous to the SPEED-improved Rama distribution, we have also tested an energy function that is averaged over the MSA in order to incorporate additional sequence information, this time via sequence correlations in the long-range interactions. The new energy function uses the original statistical potential and the same pairwise distances, $D_{i,j}$, between the pairs of amino acids. However, the new energy for each (i,j) residue pair now is the average energy calculated using the distance $D_{i,j}$ and statistical potential appropriate for the amino acid pair found in each sequence in the MSA. This procedure includes extra long-range information while maintaining the pairwise amino acid correlations inherent in each aligned sequence.

Although this method is intellectually appealing, the results are variable. We suspect that for each interaction, the optimal (lowest energy) separation distance for each contact varies too much for the different combination of residues found in the sequences in the MSA.

Consequently, the energy surface averaged across the sequences in the MSA has a shallower minimum compared to the energy function calculated using only the target sequence.

Cursory tests using a single consensus sequence with the standard energy function also fail to produce uniformly superior results. Nevertheless, we maintain that a careful and clever implementation or extension of these ideas could yield strong improvements.

Clustering. The enhancement of the fraction of native-like models obtained using SPEED has additional implications for 3° structure prediction. In our previous homology-free study, the predicted structure is the lowest energy model from the final folding ensemble. But, that structure is native-like (< 5 Å) only for about half of the targets, failing mostly when few or no accurate models are generated. Although the use of SPEED increases the proportion of accurate models, energy alone is insufficient for reliably choosing the best model. This situation is common in structure prediction. As a result, clustering methods are frequently employed because repeatedly occurring low energy conformations are typically more accurate than structurally isolated low-energy models.(Zhang and Skolnick 2004)

The lowest energy model from the top cluster for the homology-free and SPEED-based Rama distributions are presented when a cluster exists (Table 2). A larger fraction (8/12) of the SPEED-based ensembles contains identifiable clusters compared to the homology-free ensembles (6/12), and their size often is larger as well (Fig. 3). The largest cluster may be the most accurate in terms of C_α-RMSD to the native, but it may share a similar average contact profile to other less accurate clusters (Fig. 4). Most noticeable are the contact profiles of the largest two clusters of 1b72, which display almost identical contacts, but decidedly different values for the average C_α-RMSD to the native (Cluster 1, < 4 Å, Cluster 2, > 10 Å). This result is

due to the simplicity of the 1b72 fold (3-helix bundle), which permits a low energy fold that is a pseudo-mirror image fold of the native and therefore has similar contacts and similar average energy. Given this energetic similarity, the Rama distribution determines the favorability of the native conformation, with the SPEED protocol succeeding to a greater extent than the homology-free protocol.

Confidence assessed from reproducibility. While numerous methods exist for structure prediction(Skolnick et al. 2001; Srinivasan et al. 2004; Bradley et al. 2005; Ozkan et al. 2007; Yang et al. 2007; Raman et al. 2009), the quantification of the accuracy and confidence of a prediction is a crucial, but often elusive component. Template-based methods typically infer confidence from the quality of the available information used to generate an alignment and a consensus of aligned models(McGuffin 2007; Randall and Baldi 2008; Zhou and Skolnick 2008). When predicting remote templates, this technique can suffer from a dearth of PDB templates that independently align to the target sequence with high confidence. This situation precludes any meaningful clustering analysis and therefore imparts a large uncertainty to model quality.

Template-free prediction methods have an advantage of generating a large number of models that can be clustered. One noticeable feature of our method is the high correlation (R^2 = 0.85) between the average C_α-RMSD between models in the predicted cluster and the average accuracy (C_α-RMSD to the native) of the models within the cluster (Fig. 5). This trend suggests that template-free models that are reproduced with a high degree of structural similarity tend to be proportionately more accurate than models that are structurally further removed from their closest neighbors. Noticeably, the average C_α-RMSD between models in a cluster is typically

one to two angstroms lower than the average C_α-RMSD to the native of the cluster, suggesting that the top cluster has converged upon a stable but slightly non-native energy minimum. Nonetheless, this difference can be factored in when quantifying the predicted accuracy and may be diminished by improvements in the energy function and sampling distributions.

In addition to global accuracy, the residue level RMSD at each position is calculated to quantify the confidence of the prediction for each amino acid in the protein (Fig. 6). Specifically, the uncertainty at a position between models in the cluster is highly correlated to the uncertainty at that position between the cluster and the native model, suggesting that the uncertainty at each position can be predicted.

This finding has implications for other template-free methods, which may suffer method-specific difficulties when trying to quantify the confidence of model predictions. Most template-free methods rely on large fragments from PDB models,(Bradley et al. 2005; Raman et al. 2009; Zhou and Skolnick 2009) but the number of these fragments are limited and may introduce some bias due to the highly-restricted nature of their conformational search. In other words, independently converging on very similar models may not be as meaningful when the likelihood of sampling the same conformation is very high. Since the conformational changes in ItFix feature the rotation of only a single pair of φ,ψ angles, a resulting ensemble consisting of a cluster of very similar models can be treated with higher confidence given that the accessible conformation space is much larger than in fragment based methods. Similarly, the bias likely is even weaker for all-atom physics-based simulations(Ozkan et al. 2007) and *ab initio* folding simulations,(Yang et al. 2007) which have the least restricted conformational search. ItFix-SPEED may combine the best of both a restricted and unbiased conformational search in regards to assessing accuracy from the structural diversity of the largest cluster.

Performance in CASP8. We have applied an early version of the ItFix-SPEED protocol in the 2008 Critical Assessment of Structure Prediction (CASP8) for the human/server targets when a suitable template from the PDB could not be identified by the threading program RAPTOR,(Xu et al. 2003; Zhao et al. 2008) one of the top performing entries in the server category. Of these targets, the 120 residue T0482 is the only small, globular, single-domain free-modeling target with no confident templates, making it a prime candidate for the ItFix-SPEED methodology. This target has been subjected to multiple rounds of ItFix-SPEED, and our final three submitted models are very similar with highly accurate 2° and 3° structures (Fig. 7a). Our predicted 2° structure is slightly improved over the PSIPRED(Jones 1999) prediction. Due to time constraints, we initially assigned PSIPRED's high confidence (>90%) predictions at ~ 10% of the positions (total wall clock time for prediction was under 12 hours from start of prediction to submission). When the central 100 residues (ignoring the solvent exposed ends of the NMR structure) of these models are aligned to the now published structure, the C_α-RMSD to native is 4.8 Å. Hence, our algorithm is able to confidently predict the correct structure without any false positive submissions. In addition, our top model has the lowest C_α-RMSD among all submitted #1 models. We have performed commendably for other challenging template-free modeling targets, such as the D1 subdomain of protein T0405 (Fig. 6b). These results constitute strong evidence of the predictive capabilities of the ItFix-SPEED algorithm.

Our participation in CASP8 also includes predictions for sequences that have only poor templates and are considered template-free modeling targets. For target T0429, RAPTOR chooses multiple homology-based templates, but it is uncertain as to which template is correct for the C-terminal domain. ItFix-SPEED folding simulations for this domain have been used to

compare the average contact matrix of our folding simulations to the contacts of each possible template (Fig. 7d). This process has enabled us to choose a better template (T0429-2ckk) than RAPTOR's top scoring template.

The SPEED-based sampling protocol also has been used to determine the structure of the insertions of unknown structure that are present in RAPTOR-generated models. These situations have been treated by breaking the chain at one end of the insertion and then folding this free end in the context of the entire protein. The most successful outcome is for a 24 residue insertion for target T0464, where our prediction ranks as one of the top submissions (Fig. 6c).

Discussion and Conclusions

Our computationally rapid algorithm using only single (ϕ,ψ) dihedral angle moves can generate very accurate predictions of both 2° and 3° structures without relying on any known structures, templates, or fragments. For the test set, we typically predict 2° structure with ~90% accuracy, while the best 3° structure for 4/12 of the targets have C_α-RMSD below 2 Å. Hence, given intelligent search strategies and scoring functions, C_β representations can be used to accurately predict 2° and 3° structures.

Structure prediction is beyond current capabilities for the vast majority of the families identified by large-scale sequencing efforts.(Yooseph et al. 2007; Levitt 2009) The number of sequences with minimal sequence similarity to known structures is increasing at a rate that outpaces our ability to identify new families.(Levitt 2009) Currently, only about one third of the single domain architectures have known folds.(Levitt 2009)

The ItFix-SPEED procedure is well suited to contribute to mapping the protein universe, particularly for low homology sequences. Because our procedure utilizes only multiple sequence

alignments, it can take advantage of the 10^7 known sequences, and not be limited by the ~10^4 unique structures in the PDB. For CASP8 target T0482, no member of its family had a known structure, although its fold is not new. The ItFix-SPEED procedure accurately predicted its structure using only 50 non-redundant sequence homologues and no structural information. Furthermore, the ItFix-SPEED procedure is able to quantify the global and local accuracy of its prediction from the reproducibility of the trajectories, a highly desirable feature from the perspective of users of any sequence database annotation.

METHODS

Generation of Sequence Alignments. Sequence alignments are generated by PSI-BLAST(Altschul et al. 1997) using the executables from NCBI on the non-redundant database. An inter-sequence similarity cutoff of 65% is imposed with CD-HIT (Li and Godzik 2006). PSI-BLAST searches are performed in three passes with an E-value cutoff of 1.0. We choose only sequences that cover over 90% of the target sequence length and have gaps that span at most one position. These constraints are chosen such that sequences are very likely to approximate the same structure as the target. As a result of these constraints, the average E-value of each sequence in an alignment is orders of magnitude lower than 1.0.

SPEED sampling. The MSA is used to generate an amino acid substitution matrix at each position in the target sequence. Any amino acid that occurs in more than 10% of the alignments is included at that position. If a position only has only one amino acid in its substitution matrix, the amino acid occurrence threshold is decremented by 1% until there is more than 1 substitution, with the exception of proline, which is kept as the sole amino acid at a position down to 5% probability as long as there are no neighboring positions with prolines that occur at a

greater probability. If proline is the sole amino acid in the MSA-generated substitution matrix, we mutate the target sequence at that position to proline. In all other cases the sequence used during folding remains the same as the target sequence.

We initially tried calculating the SPEED distribution of a position by adding the Rama distributions at that position for each sequence in the alignment. The SPEED distributions created from this method, however, are more similar to the homology-free distribution because the target sequence amino acid usually has the highest-probability in the alignment and would be weighted proportionately in the SPEED distribution. Using a substitution matrix, on the other hand, weights all amino acids above a threshold equally, thereby rendering the resulting Rama distribution less similar to the homology-free distribution.

Since the statistics for the distributions constructed from an MSA permit many different combinations of amino acids, the area of the Rama map with vanishing probability tends to be much lower for the SPEED distribution than previously used because of the added MSA-identified combinations. In fact, the average number of angles per position used to generate a SPEED distribution is three to five-fold larger than the number of angles used to generate a homology-free distribution (Table 1). As seen in Figure 1b and the subsequent predictions, this added diversity does not dilute the specificity of the conformational search; indeed the distributions are more native-like.

Ramachandran sampling. Our prior treatment employs a sampling of specific φ,ψ angle pairs from a library generated from high resolution crystal structures, conditional on the 2° structure and nearest neighbor amino acid identities. The present study likewise employs a distribution of φ,ψ angles with the same dependencies, but instead of sampling from a large list of angles extracted from PDB models, the φ,ψ angles are chosen from a Rama distribution that is

generated for each position based on the amino acid identity and the 2° structure specification of that position and of its nearest neighbors. Thus, Rama distributions are calculated for the central residue in each of the distinct 8000 combinations of three contiguous amino acids, conditional upon the amino acid identity and on the 2° structure of all three residues. Because the ItFix simulations consider six possible categories of 2° structure for the construction of the sampling distributions (H: helix, E: strand, C: coil, A: everything, O: not helix, Q: Not strand), 1,728,000 possible Rama distributions are constructed to describe the possible 8000 amino acid triplets. Each Rama distribution has 72^2 5°x5° bins, and each bin is assigned a probability that is determined by frequency of occurrence of these backbone dihedral angles in the PDB for the specific conditions of amino acid identities and 2^o structure. A Rama distribution accommodates the increase in PDB-derived angles introduced by SPEED without increasing the system memory, as occurs when each angle is explicitly stored in memory.

The sampling of φ,ψ angles begins by selecting a bin in Rama space according to the probability assigned to that bin (e.g., a bin that contains 1.5% of the angle counts for the distribution at that position has a 0.015 probability of being selected). This bin selection is followed by the selection of a random angle uniformly from within the 5°x5° window of that bin. The Rama distribution of the central residue of the triplet INE (position 4 in 1tif) with all allowed 2° structures is an example of one such sampling distribution (Fig 1b, top). If the subsequent round of ItFix eliminates a 2° structure option at a position, the Rama distribution at that position is changed accordingly (Fig. 1b, middle, bottom).

Clustering algorithm. After the ItFix protocol generates a predicted 2^o structure, a further 10,000 folding simulations are run to maximize the exploration of conformational space. The

pairwise C_α-RMSD matrix of the resulting 10,000 models is used to cluster the ensemble into groups of models that all align to each other below a C_α-RMSD cutoff, an approach that is similar to the SPICKER algorithm (Zhang and Skolnick 2004). Other methods (Gong et al. 2005) cluster according to the C_α-C_α distance instead of the pairwise C_α-RMSD, but we find that the C_α-C_α distances in some cases are highly correlated even though the C_α-RMSD between the models are quite different (Fig. 4b).

When identifying clusters, the upper limit of the cutoff distance of the inter-model C_α-RMSD is increased in increments of 1 Å starting at 1 Å until at least five clusters are found, or a 7 Å limit is reached. Every model in the cluster must have a C_α-RMSD to every other model in the cluster that is less than the cutoff distance. Targets with predicted all-α 2° structures have a minimum cluster size of 5%, whereas the minimum size for targets with other predicted 2° structure types can be as low as 0.04%. A cluster is eliminated if it contains a model present in a larger cluster. The largest cluster is selected as the predicted model, unless it has an above average energy and there is another cluster with an energy that is greater than one standard deviation below average.

Model refinement. One of the most important challenges of structure prediction is an effective exploration of conformational space. Ideally an exhaustive refinement is performed for every model generated by folding, but we take a computationally thrifty approach and refine only the models in the largest cluster of each target. Refinement consists of the same move set and energy function as folding, with the addition of the fact that we reject moves that increase the *Rg*, *Br* or *Ru* of the starting model. Each model in the cluster is refined 100 times and the model with the lowest average energy among all the refined models is chosen at the prediction listed in Table I.

Parallel scripting with Swift. The ItFix-SPEED algorithm has been implemented, tested and evaluated(Hocky 2009) using an innovative parallel scripting language called Swift(M. Wilde 2009). The Swift runtime system automates parallelization, data management, and error recovery, and supports execution on a wide variety of parallel computer systems. This allows the composition of flexible structure prediction scripts to address new energy functions and explore algorithm enhancements, and to compare the behavior of the algorithm under a wide range of conditions and parameter settings.

Electronic supplementary material. Three supplementary figures and one table.

Figure S1. Position-based comparison of homology-free and SPEED distributions. The analysis of Figure 2c is shown for additional targets.

Figure S2. Reproducibility of folding ensemble. The folding ensembles of 10,000 models are divided into five sets of 2000 models for the targets 1dcj, 1tif, and 1r69. The lack of diversity illustrates this evaluation metric is highly reproducible.

Figure S3. Effect of energy filtering on prediction accuracy. The accuracy of the folding ensemble increases as higher energy models are removed. Shown is the fraction of models below varying C_α-RMSD cutoffs. Traces represent the results after removal of models with energies higher than E greater than <Energy>+X where X=0,±σ, ±2σ, and σ is the standard deviation in energy for all models.

Supplementary Table S1. Radial values for targets

Acknowledgments. We thank members of the Sosnick and Freed labs for helpful conversation. Many of the simulations in this work were run under the Swift scripting framework, and we thank the Swift and Falcon developers for their support. This work was supported by National Institutes of Health research (to T.R.S., K.F.F., J.X.) and training grants (to J.D.), National Science Foundation grants OCI-721939 and OCI-0944332 (to M.W.), TeraGrid resources provided by the National Center for Supercomputing Applications, The LSU Center for Computing Technology, the Texas Advanced Computing Center, by the Argonne Leadership Computing Facility, and the U.S. Department of Energy under Contract DE-AC02-06CH11357.

References.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389-3402.

Baker, D., and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294:** 93-96.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res* **30:** 276-280.

Bradley, P., Misura, K.M., and Baker, D. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science* **309:** 1868-1871.

Bryson, K., McGuffin, L.J., Marsden, R.L., Ward, J.J., Sodhi, J.S., and Jones, D.T. 2005. Protein structure prediction servers at University College London. *Nucleic Acids Res* **33:** W36-38.

Cheng, J., Randall, A.Z., Sweredoski, M.J., and Baldi, P. 2005. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* **33:** W72-76.

DeBartolo, J., Colubri, A., Jha, A.K., Fitzgerald, J.E., Freed, K.F., and Sosnick, T.R. 2009. Mimicking the folding pathway to improve homology-free protein structure prediction. *Proc Natl Acad Sci U S A* **106:** 3734-3739.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* **19:** 99-113.

Gong, H., Fleming, P.J., and Rose, G.D. 2005. Building native protein conformation from highly approximate backbone torsion angles. *Proc Natl Acad Sci U S A* **102:** 16227-16232.

Hocky, G., Wilde, M., Debartolo, J., Hategan, M., Foster, I., Sosnick, T.R., and Freed, K.F. 2009. Homology-free protein structure prediction through parallel scripting. *Argonne Technical Report Preprint ANL/MCS-P1645-0609.* .

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292:** 195-202.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577-2637.

Kortemme, T., Morozov, A.V., and Baker, D. 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326:** 1239-1259.

Krantz, B.A., Dothager, R.S., and Sosnick, T.R. 2004. Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. *J. Mol. Biol.* **337:** 463-475.

Levitt, M. 2009. Nature of the protein universe. *Proc Natl Acad Sci U S A* **106:** 11079-11084.

Li, W., and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658-1659.

Li, W., Jaroszewski, L., and Godzik, A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17:** 282-283.

M. Wilde, I.F., K. Iskra, P. Beckman, Z. Zhang, A. Espinosa, M. Hategan, B. Clifford, I. Raicu. 2009. Parallel scripting for applications at the petascale and beyond. *IEEE COMPUTER*.

McGuffin, L.J. 2007. Benchmarking consensus model quality assessment for protein fold recognition. *BMC Bioinformatics* **8:** 345.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247:** 536-540.

Ozkan, S.B., Wu, G.A., Chodera, J.D., and Dill, K.A. 2007. Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. U S A* **104:** 11987–11992.

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47:** 228-235.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., Dimaio, F., Lange, O., et al. 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **20:** 20.

Randall, A., and Baldi, P. 2008. SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. *BMC Struct Biol* **8:** 52.

Skolnick, J., Fetrow, J.S., and Kolinski, A. 2000. Structural genomics and its importance for gene function analysis. *Nat Biotechnol* **18:** 283-287.

Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P., and Boniecki, M. 2001. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins* **Suppl 5:** 149-156.

Soding, J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21:** 951-960.

Sosnick, T.R., Krantz, B.A., Dothager, R.S., and Baxa, M. 2006. Characterizing the Protein Folding Transition State Using psi Analysis. *Chem. Rev.* **106:** 1862-1876.

Srinivasan, R., Fleming, P.J., and Rose, G.D. 2004. Ab initio protein folding using LINUS. *Methods Enzymol* **383:** 48-66.

Xu, J., Li, M., Kim, D., and Xu, Y. 2003. RAPTOR: optimal protein threading by linear programming. *J. Bioinform. Comp. Biol.* **1:** 95-117.

Yang, J.S., Chen, W.W., Skolnick, J., and Shakhnovich, E.I. 2007. All-atom ab initio folding of a diverse set of proteins. *Structure* **15:** 53-63.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., et al. 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5:** e16.

Zhang, Y., and Skolnick, J. 2004. SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* **25:** 865-871.

Zhao, F., Li, S., Sterner, B.W., and Xu, J. 2008. Discriminative learning for protein conformation sampling. *Proteins* **73:** 228-240.

Zhou, H., and Skolnick, J. 2008. Protein model quality assessment prediction by combining fragment comparisons and a consensus C(alpha) contact potential. *Proteins* **71:** 1211-1218.

Zhou, H., and Skolnick, J. 2009. Protein structure prediction by pro-Sp3-TASSER. *Biophys J* **96:** 2119-2127.

Table 1 **SPEED 2° structure prediction comparison**[1]

Protein				Rama Enrichment[2] (angles / residue)		2° structure accuracy[3] Q3 (Q8)			
PDB ID	size	fold	$N_{E}{}_{FF}^{4}$	Hfree	SPEED	ItFix	ItFix SPEED	SSPro	PSI-PRED
1af7	69	α	7.3	1426	5599	97 (86)	96 (88)	86 (81)	90
1b72	50	α	5.7	1384	4229	88 (84)	96 (96)	68 (72)	84
1csp	67	β	6.0	1069	2365	79 (67)	87 (70)	75 (67)	88
1di2	68	αβ	6.8	1230	4964	88 (79)	85 (74)	74 (75)	97
1dcj	72	αβ	7.0	1059	4381	45 (29)	83 (65)	65 (56)	89
1mky	77	αβ	5.0	1572	3947	86 (70)	83 (65)	87 (71)	90
1o2f	77	αβ	5.5	1059	4506	78 (69)	84 (73)	79 (66)	75
1r69	61	α	7.5	1036	5058	93 (89)	97 (89)	74 (72)	92
1shf	59	β	7.1	774	3213	76 (56)	71 (51)	85 (69)	80
1tif	57	αβ	4.4	1349	3233	89 (79)	91 (81)	76 (70)	93
1tig[5]	86	αβ	5.4	1194	3323	83 (70)	N/A	69 (67)	83
1ubq	73	αβ	7.7	1152	3405	92 (69)	94 (82)	88 (67)	90

[1]Target sequences are from our previous homology-free ItFix study,(DeBartolo et al. 2009) which have been selected from a previous Rosetta prediction study.(Bradley et al. 2005).

[2]Rama enrichment is the positional average of the number of PDB angles used to generate the Rama distribution for each method. The Q3 and Q8 (in parentheses) 2° structure prediction

accuracies are reported for the previous homology-free study, an updated homology-free version, and SPEED sampling.

[3]SSpro and PSIPRED 2° structure predictions are obtained from their respective servers.(Bryson et al. 2005; Cheng et al. 2005)

[4]N_{EFF}(Soding 2005) is a Shannon entropy measure on a scale of 1-20 of the amino acid diversity of the sequence alignment (1 = single amino acid, 20 = all amino acids are equally likely).

[5]Folding of 1tig could not converge in reasonable amount of time because radial terms could not be satisfied in a small number of MCSA steps.

Table 2. **3° structure prediction**

Protein				3° structure accuracy			
PDB ID	size	fold	N_{EFF}	Previous ItFix[1]	ItFix-hfree[2]	ItFix-SPEED[3]	C_α-5.0X[4]
1af7	69	α	7.3	2.9 (2.5)	2.5 (2.5)	2.6 (1.6)	1.2
1b72	50	α	5.7	3.5 (1.6)	3.6 (1.7)	3.5 (1.6)	1.1
1csp	67	β	6.0	10.5 (6.0)	NC (4.6)	5.2 (4.1)	4.2
1di2	68	αβ	6.8	6.1 (4.6)	10.2 (6.1)	9.6 (6.7)	N/A
1dcj	72	αβ	7.0	13.3 (7.6)	NC (5.9)	5.3 (4.6)	∞
1mky	77	αβ	5.0	6.9 (6.1)	NC (4.4)	5.2 (4.2)	∞
1o2f	77	αβ	5.5	11.2 (5.8)	NC (6.7)	NC (4.2)	∞
1r69	61	α	7.5	4.2 (2.4)	3.7 (2.1)	3.5 (1.6)	1.8
1shf	59	β	7.1	12.2 (6.7)	NC (6.2)	NC (3.8)	∞
1tif	57	αβ	4.4	11.3 (4.2)	5.7 (3.7)	5.4 (3.2)	4.3
1tig[5]	86	αβ	5.4	6.4 (5.3)	N/A	N/A	N/A
1ubq	73	αβ	7.7	5.3 (3.1)	4.4 (3.6)	2.6 (1.9)	6.0

[1]The C_α-RMSD to the native of prediction based on energy and best model (in parentheses) from our previous homology-free ItFix study.(DeBartolo et al. 2009)

[2]Folding with the homology-free Rama distribution and with the final SPEED 2° structure (2000 trajectories), cluster and refinement prediction and best model (in parentheses).

[3]Folding with the SPEED Rama distribution with final SPEED 2° structure (10,000 trajectories), cluster and refinement prediction and best model (in parentheses).

[4]Ratio of the percentage of models below 5.0 Å C_α-RMSD to native of SPEED (column 7) to homology-free (column 6)

[5]Folding of 1tig could not converge in reasonable amount of time because radial terms could not be satisfied in a small number of MCSA steps.

Figure legends

Figure 1. Structure prediction protocol. a) The 2° and 3° structure prediction protocol for homology-free modeling uses the target sequence to generate a Rama sampling distribution, whereas SPEED uses a distribution that is averaged over a Multiple Sequence Alignment (MSA). The ItFix algorithm iteratively defines the 2° structure, and clustering and refinement are used to predict 3° structure. **b)** The Rama distribution for position 4 of the sequence of 1tif is shown for representative rounds of ItFix for homology-free and SPEED sampling. The native φ,ψ angles are denoted as a red circle.

Figure 2. SPEED-enhanced ϕ,ψ sampling distribution. a) Rama space is divided into four coarse regions for analysis. **b)** The percentage of residues with probability exceeding 0.25 for the native Rama region is increased for SPEED for all targets, particularly 1csp and 1dcj. **c)** For 1b72, the probability of the native Rama region is greatly enhanced using SPEED .

Figure 3. Improvement in 3° structure prediction using SPEED. The percentage of models with a C_α-RMSD to the native below a cutoff level (x-axis) provides a comparison of the overall accuracy of the folding ensembles. The top cluster (solid line) from SPEED is much better than the entire SPEED ensemble (dashed line), which is better than the ensemble generated using the homology-free ItFix Rama distribution with the SPEED-generated 2° structure assignments (dotted line).

Figure 4. Comparison of contacts for the top clusters of several targets. Each map is a C_α-C_α contact matrix with a 10.0 Å distance cutoff for α targets 1af7, 1b72 and a 8.0 Å distance cutoff for the αβ, β targets 1mky, and 1csp. Contacts of the native model are presented on the lower right of each map. The largest cluster for 1af7 has the most native contacts and has an average C_α-RMSD to the native below 4 Å. The next largest 1af7 cluster, which has an averge greater than 10 Å C_α-RMSD to the native, exhibits many native and non-native contacts. The largest 1b72 cluster is the most native in terms of C_α-RMSD (< 3Å average), but contains identical contacts to the next largest cluster (> 10 Å C_α-RMSD to native average) that is the mirror-image fold of the native. The contacts matrices of the largest clusters of 1mky and 1csp are both very native-like.

Figure 5. Assessing global accuracy from reproducibility of the top cluster. The mean C_α-RMSD to native of the top cluster is strongly correlated with the C_α-RMSD between the models in that cluster, indicating that the latter metric can be used as a measure of predicted model's accuracy.

Figure 6. Assessing local accuracy from reproducibility of top cluster. Position-resolved model accuracy and confidence. The average aligned distance between all models in the predicted cluster is determined for each position. This value is highly correlated to the average aligned distance at each position between each model in the cluster and the native structure. The standard deviation for each of these values also is highly correlated, suggesting the ability to use clustering to determine confidence for each position in a predicted model.

Figure 7. ItFix-SPEED blind predictions in CASP8. a) 2° and 3° structure prediction of target T0482. The ItFix 2° structure prediction compares favorably to the native at 84% accuracy, which is slightly superior to the 82% accuracy of PSIPRED. The Global Distance Test (GDT) value is the % of the residues within a cut-off distance of the native structure. This cut-off distance is the y-value on the plot (e.g., for the ItFix prediction, 83% and 100% of the residues are predicted to within 4.7 and 7.8 Å of the native structure, respectively). The GDT trace for our ItFix prediction (blue line) is the rightmost of all the Model 1 predictions indicating that the methods is able to predict more residues with higher accuracy. The Itfix-SPEED prediction for **b)** the entire Domain 1 of target T0405, and **c)** the 24 residue insertion in RAPTOR's predicted template for T0464. **d)** Itfix-SPEED selection of the best template identified by RAPTOR based on average predicted tertiary contacts. Contact map, upper left: ItFix average contacts for the final structures from 100+ folding trajectories; lower right: contacts of one of RAPTOR's lower ranked templates, which is the closer to the native structure than its top ranked template, which has a less similar contact map. Values in parenthesis are the C_α-RMSD between predictions and the native structure. GDT plots are taken from the CASP8 website (www.predictioncenter.org/casp8/index.cgi).